Nonlinear methods in solving ordinary differential equations

A. Wambecq (*)

ABSTRACT

Some one step methods, based on nonpolynomial approximations, for solving ordinary differential equations are derived, and numerically tested. A comparison is made with existing methods.

0. INTRODUCTION

Consider the problem of finding a numerical solution for the ordinary differential equation :

$$\begin{cases} \dfrac{dy}{dx} = f(x, y) \\ y(a) = 0 \end{cases} \tag{0.1}$$

in the interval [a, b]. To achieve this one can use one step methods of the form :

$$y_{n+1} = g(x_n, y_n, h) \tag{0.2}$$

where $x_n = a + (n-1)h$, $n = 1, \ldots, N$, h being a fixed step size, y_n an approximate value for $y(x_n)$, and $N = (b-a)/h$.
A one step method of the form (0.2) can be derived by choosing $g(x, y, h)$ such that :

$$g(x, y, h) - s(x, y, h) = 0(h^{p+1}),$$

where

$$s(x, y, h) = y + hf(x, y) + \frac{h^2}{2!} f'(x, y) + \ldots .$$

In such case, the one step method is said to be of order p.
A simple way to achieve this is to take $g(x, y, h)$ equal to a partial sum of $s(x, y, h)$. The corresponding one step method :

$$y_{i+1} = y_i + hf_i^{(0)} + \frac{h^2}{2!} f_i^{(1)} + \ldots + \frac{h^p}{p!} f_i^{(p-1)} \tag{0.3}$$

is called the Truncated Taylor Series (TTS) method of order p. (Lambert 1974, p. 46). Here $f_i^{(j)}$ is a short notation for $f^{(j)}(x_i, y_i)$.

The aim of this paper is to derive one step methods where g is a nonlinear function of f and its derivatives. In section 1, some nonlinear methods derived from the TTS method are given together with an efficient way of using such methods. These methods are numerically compared with the TTS methods in section 2. In section 3 nonlinear Runge-Kutta type methods are derived. Some properties of these methods will be given.

1. NONLINEAR METHODS DERIVED FROM TRUNCATED TAYLOR SERIES

A variety of nonlinear methods can be derived by using Padé approximations for s. We first give the definition of the Padé approximant (Padé 1892; Gragg 1972) for the power series :

$$C(x) = c_0 + c_1 x + c_2 x^2 + \ldots$$

Let R_m^ℓ denote the class of rational functions $r = \frac{p}{q}$, where p respectively q is a polynomial of degree at most ℓ respectively in, such that $\frac{p}{q}$ is irreducible.
The (ℓ, m)th Padé approximant is the unique element $r_{\ell, m} = \frac{p}{q}$ in R_m^ℓ such that

$$p(x) - q(x) C(x) = o(x^{\ell + m + 1 + k}) \tag{1.1}$$

for some integer value of k which is as high as possible. The elements $r_{\ell, m}$ for $\ell, m \geqslant 0$ can be arranged in a table as follows :

$r_{0,0}$	$r_{1,0}$	$r_{2,0}$	$r_{3,0}$	$\cdots\cdots$
$r_{0,1}$	$r_{1,1}$	$r_{2,1}$	$r_{3,1}$	$\cdots\cdots$
$r_{0,2}$	$r_{1,2}$	$r_{2,2}$	$r_{3,2}$	$\cdots\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

(*) A. Wambecq, Applied Mathematics and Programming Division, K. U. Leuven, Celestijnenlaan 200 B, B-3030 Heverlee, Belgium

This table is called the Padé table for C(x). Remark that the elements on the first row are equal to the partial sums of C(x).
A Padé table is said to be normal if no two elements of the table are equal (Padé 1892).
All Padé tables in this paper will be considered to be normal.

a) The use of the Padé table for s

A special case of the methods that can be derived from the Padé table for s was already given by Lambert and Shaw (1965).
They proposed to use a rational interpolant of the form :

$$F(x) = \frac{\sum_{i=0}^{p} a_i x^i}{b_0 + x},$$

subject to the conditions :

$$\begin{cases} F^s(x_n) = f^{(s-1)}(x_n, y_n) \quad s = 1, \ldots, p+1 \\ F(x_n) = y_n \\ F(x_n + h) = y_{n+1} \end{cases}$$

By eliminating the $p+2$ coefficients $a_0, \ldots, a_p, b_0$ from these $p+3$ conditions, one gets a one step method of the form :

$$y_{n+1} = y_n + \sum_{i=1}^{p-1} \frac{h^i}{i!} f_n^{(i-1)} + \frac{h^p}{p!} \frac{(p+1)(f_n^{(p-1)})^2}{(p+1) f_n^{(p-1)} - h f_n^{(p)}} \tag{1.2}$$

Now consider the following power series in h :

$$s(h) = s(x_n, y_n, h) = y_n + h f(x_n, y_n) + \frac{h^2}{2!} f'(x_n, y_n) + \ldots$$

Let r be the (p, 1)-th Padé approximant of s.
Putting $y_{n+1} = r(x_{n+1})$, we get

$$y_{n+1} = y_n + \sum_{i=1}^{p} \frac{h^i}{i!} f_n^{(i-1)} + \frac{h^p}{p!} \frac{h f_n^{(p)} f_n^{(p-1)}}{(p+1) f_n^{(p-1)} - h f_n^{(p)}} \tag{1.3}$$

Taking together the terms in h^p in formula (1.3), one finally gets formula (1.2).
More general, the following theorem can be proved:

Theorem

The one step method derived by using the (ℓ, m)th Padé approximant of s is the same as the method derived by eliminating the coefficients a_i, b_j from :

$$\begin{cases} F(x) = \dfrac{\sum_{i=0}^{\ell} a_i x^i}{\sum_{j=0}^{m} b_j x^j} \\ F^{(s)}(x_n) = f^{(s-1)}(x_n, y_n), \quad s = 1, \ldots, \ell + m \\ F(x_n) = y_n \\ F(x_n + h) = y_{n+1} \end{cases}$$

b) Some properties of the Padé table for s (s table)

Let $t(x, y, h) = \dfrac{s(x, y, h) - y}{h}$ (1.4)

or $t(x, y, h) = f(x, y) + \frac{h}{2!} f'(x, y) + \frac{h^2}{3!} f''(x, y) +$

then t can also be considered as a power series in h, and therefore has a Padé table (t table).

Let $\dfrac{D_s}{N_s}$ be the (ℓ, m)th element of the s table, and

$\dfrac{D_t}{N_t}$ the $(\ell - 1, m)$th element of the t table

In the case where both s and t tables are normal we have :

$$\frac{D_t}{N_t} = t + O(h^{\ell + m}) \tag{1.5}$$

and

$$\frac{D_s}{N_s} = s + O(h^{\ell + m + 1}) \tag{1.6}$$

Let $v = y + h \dfrac{D_t}{N_t}$. (1.7)

Using (1.4) and (1.5) we get

$$v = s + O(h^{\ell + m + 1}) \tag{1.8}$$

Although the order of approximation for the series s of (1.6) and (1.8) is the same, v and $\dfrac{D_s}{N_s}$ are not always identical.
To illustrate this remark, we consider the cases $\ell \geqslant m$ and $\ell < m$ separately. For $\ell \geqslant m$, relation (1.7) implies that $v \in R_m^{\ell}$, and the normality of the s table implies that v is irreducible, and so is $\dfrac{D_s}{N_s}$.
Consequently, by uniqueness of the Padé approximation $v = \dfrac{D_s}{N_s}$.
On the contrary for $\ell < m$, $v \in R_m^m$ is not in general a Padé approximation anymore, for its order of approximation is not $2m + 1$ as it should be.

Example

Consider the case where $\ell = 1$, $m = 2$.

Put

$$q_1 = (2h^2 ff^{(2)} - 3h^2 (f^{(1)})^2 + 6hff^{(1)} - 12f^2)$$

$$q_2 = (q_1 - 2hf^{(2)} y + 6f^{(1)} y)$$

Then

$$v = y + h \frac{-12f^3}{q_1}$$

and

$$\frac{D_s}{N_s} = \frac{-2(6hf^3 - 6hff^{(1)} y + hf^{(2)} y^2 + 6f^2 y - 3f^{(1)} y^2)}{q_2}$$

The difference between v and $\frac{D_s}{N_s}$ is given by

$$v - \frac{D_s}{N_s} = \frac{-h^4 y(4f^2 (f^{(2)})^2 - 12f(f^{(1)})^2 f^{(2)} + 9(f^{(1)})^4)}{q_1 q_2}$$

Both v and $\frac{D_s}{N_s}$ give an approximation to s up to the term h^3. Hence their difference is of order h^4, and in general not identical zero.

In a similar way as for the s table, nonlinear methods for solving (0.1) can be derived from the t table. These methods are then of the form

$$y_{n+1} = y_n + hr(x_n, y_n, h),$$

where r is a Padé approximant for $t(x_n, y_n, h)$ of a certain order.

The conclusion of this section is that if $\ell \geqslant m$, there is no profit in using the s table for deriving a method of the form (0.2), since by using the t table the amount of computational work can be decreased.

c) Practice

In practice, one can get great flexibility in choosing the type of rational interpolant simply by using a suitable algorithm for computing the elements of the Padé table for t (e. g. the ϵ algorithm, Wynn 1962).

The use of these algorithms in numerical work also eliminates the cumbersone derivation of unwieldy rules like (1.2). As a consequence also Padé approximants of order (ℓ, m) with $m > 1$ can be used in practice, which seemed to be very complicated if explicit expressions for g in (0.2) are used. (See the remark of Lambert & Shaw 1965).

Notice further that methods, derived from the (ℓ, m)th element of the s table, produce the exact values of the solutions of (0.1), if this solution is a function of the class R_m^{ℓ}.

2. NUMERICAL RESULTS

The rational methods were tested in various cases, using the ϵ algorithm for computing the values of the t table.

Only in a few cases some of the rational methods gave inaccurate solutions. See for example problem f) in the table below. The inaccuracy was due to a breakdown of the ϵ algorithm in the first step, caused by the zero of the solution $y = tg(x)$ at the origin. Nevertheless, the results of the (ℓ, ℓ)th elements were quite acceptable, and still better than the results obtained by the TTS methods of the same order.

In most examples the results obtained by using the approximants near the main diagonal in the table were extremely better than the corresponding results for the TTS method.

The seven numerical problems below were run on an IBM 370/158 computer, using double precision.

	equation	interval [a, b]	solution
a)	$y' = y$, $y(0) = 1$	[0,3]	$y = e^x$
b)	$y' = -y$, $y(0) = 1$	[0,6]	$y = e^{-x}$
c)	$y' = x + y$, $y(0) = 1$	[0,1]	$y = 2e^x - x - 1$
d)	$y' = 2(x+1)y$, $y(0) = e$	[0,1]	$y = e^{(x+1)^2}$
e)	$y' = 1 + y^2$, $y(0) = 1$	[0,0.76]	$y = tg(x + \frac{\pi}{4})$
f)	$y' = 1 + y^2$, $y(0) = 0$	[0,1.4]	$y = tg(x)$

Stepsize : $h = [b-a]/50$

In the following tables, the relative error for computing y(b) in the above six examples is given. The relative error is defined by :

$$e_r = \left| \frac{y(b) - y_N}{y(b)} \right| \text{, with } N = 50.$$

The errors are rounded to two significant figures and the exponents of these numbers are enclosed within brackets.

Problem a)

m \ ℓ	0	1	2	3	4
0	0.12(+00)	0.40(-02)	0.87(-03)	0.20(-05)	0.31(-07)
1	0.21(-02)	0.36(-04)	0.43(-06)	0.80(-08)	
2	0.32(-06)	0.30(-06)	0.34(-08)		
3	0.11(-02)	0.25(-06)			
4	0.42(+00)				

Problem b)

m\ℓ	0	1	2	3	4
0	0.40(+00)	0.41(-01)	0.26(-02)	0.84(-04)	0.38(-05)
1	0.28(-01)	0.80(-03)	0.24(-04)	0.44(-05)	
2	0.12(-04)	0.12(-04)	0.21(-06)		
3	0.11(-01)	0.30(-05)			
4	0.20(+02)				

Problem c)

m\ℓ	0	1	2	3	4
0	0.22(-01)	0.13(-03)	0.16(-05)	0.31(-08)	0.20(-10)
1	0.20(-03)	0.32(-06)	0.18(-08)	0.36(-11)	
2	0.26(-05)	0.35(-10)	0.29(-11)		
3	0.23(-07)	0.17(-11)			
4	0.35(-09)				

Problem d)

m\ℓ	0	1	2	3	4
0	0.14(+00)	0.62(-02)	0.10(-03)	0.47(-05)	0.79(-07)
1	0.12(-02)	0.43(-04)	0.58(-06)	0.17(-07)	
2	0,90(-05)	0.26(-06)	0.26(-08)		
3	0.25(-06)	0.84(-08)			
4	0.40(-08)				

Problem e)

m\ℓ	0	1	2	3	4
0	0.50(+00)	0.11(+00)	0.53(-01)	0.12(-01)	0.21(-02)
1	0.26(-02)	0.53(-05)	0.57(-07)	0.21(-09)	
2	0.33(-07)	0.35(-07)	0.56(-09)		
3	0.10(-02)	0.28(-07)			
4	0.10(+01)				

Problem f)

m\ℓ	0	1	2	3	4
0	0.11(+00)	0.51(−02)	0.27(−03)	0.14(−04)*	0.78(−06)
1	0.13(+01)	0.14(−04)	0.57(−05)*	0.45(−09)	
2	0.10(+00)	0.10(+00)*	0.21(−10)		
3	0.19(+01)*	0.19(+01)			
4	0.19(+01)				

The errors of the TTS methods can be found in the first row (m = 0) of each table.
If only the evaluation of f and its derivatives is counted, the computational work stays the same along the bottom left, upper right diagonal. The order of approximations also stays the same along this diagonal (see for example the elements marked with (*) in the table for example f)). Considering these facts, the accurate results of the rational methods near the main diagonal are quite striking.

3. RUNGE-KUTTA TYPE METHODS

Often the higher derivatives of f are too hard to compute. In such case, one can try to replace $f', f'', \ldots,$ by linear combinations of evaluations of f at different points.
In the linear case, this gives rise to the well known Runge-Kutta methods.

Let

$$k_j = f\left(x_i + p_j h, y_i + h \sum_{\ell=1}^{j-1} b_{j\ell} k_\ell\right) \quad \text{for } j = 1, \ldots, n.$$

Then

$$y_{i+1} = y_i + h \sum_{j=1}^{n} a_j k_j \qquad \text{for } i = 0, 1, 2, \ldots$$

The parameters a_j, $b_{j\ell}$ and p_j are choosen such that the power series in h coincides with the series t in as many terms as possible.
Instead of restricting ourself to linear combinations, we will consider nonlinear combinations of function evaluations of f.
To illustrate this technique we first give an example.
Let

$$k_1 = f(x_i, y_i)$$

$$k_2 = f(x_i + ph, y_i + phk_1) \tag{3.1}$$

Consider $y_{i+1} = y_i + h \dfrac{k_1^2}{ak_1 + bk_2}$

Here, a, b and p will be determined so that the power series expansion of the nonlinear term agrees with the power series t up to the terms of order two in h. This gives the condition :

$$(f_i + \frac{h}{2} f_i')(af_i + b(f_i + phf_i')) - f^2 = 0(h^2).$$

This equality must be true for all functions f whose first derivative exists. So the nonlinear system to solve is :

$$\begin{cases} a + b = 1 \\ bp = -\dfrac{1}{2} \end{cases} \tag{3.2}$$

Putting $b = -1$ in (3.2), the following one step method is derived :

$$k_1 = f(x_i, y_i)$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2} h k_1\right)$$

$$y_{i+1} = y_i + h \frac{k_1^2}{2k_1 - k_2}$$

This method is of order 2, and where f(x, y) is a linear function in x and y, its truncation error is half the truncation error of the classical Euler-Cauchy rule

$$y_{i+1} = y_i + hf\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f_i\right).$$

The use of the particular form of the function g (0.2) in the above example was inspired by the form of the elements of a pseudo Padé table for t, which is defined as follows.
Consider a table where the elements of the first row contain linear combinations of function evaluations of f, such that the order of approximation to t of this combination increases with increasing column number. If one formally applies the ϵ algorithm (Wynn 1956) on this row of elements, one gets nonlinear combinations in the other rows. Some elements of such a table are given below.
The derivation of nonlinear Runge-Kutta methods can be achieved in several ways. One of them is the "brute force" method. Suppose we would like to find a method of the form :

$$y_{i+1} = y_i + h \frac{F(k_1, k_2, \ldots, k_n)}{G(k_1, k_2, \ldots, k_n)}$$

where

$$k_j = f(x_i + p_j h, y_i + \sum_{\ell=1}^{j-1} a_{\ell j} k_\ell).$$

To get this method we can try to solve the nonlinear system in the coefficients of $\frac{F}{G}$ and the parameters p_j, $a_{\ell j}$, to obtain a method of as high an order as possible, i. e. :

$$\frac{F}{G} = t + 0(h^\nu).$$

A disadvantage of doing so is that not all forms $\frac{F}{G}$ are suitable to solve the problem satisfactorily. The resulting nonlinear system is often non-consistent. This is illustrated by the following example : consider

$$\frac{F}{G} = \frac{k_1}{ak_1 + bk_2},$$

with k_1 and k_2 as in (3.1). If we try to construct an order two method with this form, the following condition must be satisfied :

$$(f_i + \frac{h}{2} f_i')(af_i + b(f_i + phf_i')) - f_i = 0(h^2).$$

This equality must hold for all functions f, so that the nonlinear system for the coefficients a, b and parameter p is :

$$\begin{cases} a + b = 0 \\ bp = -1 \\ -1 = 0 \end{cases}$$

and this system is, of course, non-consistent.

An easier way of deriving nonlinear methods is to start from known linear ones, and to apply formally some convergence accelerating rule on them.
A special case of this is the pseudo Padé table mentioned above. Moreover, the elements of such a pseudo Padé table have interesting properties in connection with the Padé table for t.
Indeed, consider the (ℓ, m)th element $\frac{p}{q}$ from the Padé table for t, and the $(\ell, m)$the element $\frac{F}{G}$ from a pseudo Padé table for t.

Lemma

Let the Padé table for t be normal. Provided the $(\ell, 0)$th elements of the pseudo Padé table give approximations of order $\ell + 1$ exactly, the (ℓ, m)th element $\frac{F}{G}$ of the

Table of some nonlinear methods derived from known linear methods.

$m \backslash \ell$	0	1	2
0	k_1	$\frac{3k_2 - k_1}{2}$	$\frac{k_1 + 3k_3}{4}$
1	$\frac{2k_1^2}{5k_1 - 3k_2}$	$\frac{3k_2^2 - k_1k_3 - 2k_1k_2}{4k_2 - 3k_1 - k_3}$	
2	$\frac{8k_1^3}{16k_2^2 - 6k_1k_3 + 36k_1k_2 - 32k_1^2}$		

$$\begin{cases} k_1 = f(x_i, y_i) \\ k_2 = f(x_i + \frac{h}{3}, y_i + \frac{h}{3} k_1) \\ k_3 = f(x_i + \frac{2}{3} h, y_i + \frac{2}{3} hk_2) \end{cases}$$

pseudo Padé table gives an approximation to t of order $k = \ell + m + 1$.

Proof

The proof of this lemma is easily done by induction on ℓ and m, using the relations necessary to construct the pseudo Padé table (Wynn 1961).

Theorem

Suppose the Padé table for t and the pseudo Padé table for t meet the requirements of the above lemma. Then the Padé table for t and the Padé table for the (ℓ, m)th element $\frac{F}{G}$ of the pseudo Padé table for t are identical for all elements (i, j) for which $i+j \leqslant \ell+m+1 = k$.

Proof

By the above lemma, the order of approximation of $\frac{F}{G}$ is k, so that one can write :

$$t - \frac{F}{G} = 0(h^k) \tag{3.3}$$

Let

$$t = c_0 + c_1 h + c_2 h^2 + \ldots + c_{k-1} h^{k-1} + c_k h^k + \ldots,$$

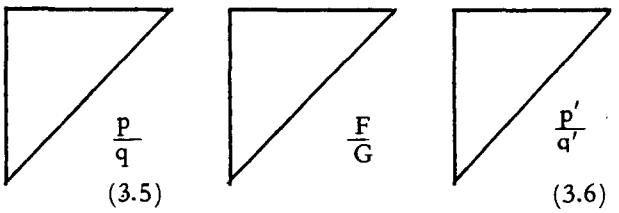
then the power expansion of $\frac{F}{G}$ is given by

$$\frac{F}{G} = c_0 + c_1 h + c_2 h^2 + \ldots + c_{k-1} h^{k-1} + 0(h^k).$$

Now the Padé approximants of order k and less are only dependent of the first k terms in the power expansion of the function. So the elements of order $p \leqslant k$ in the Padé table for t are equal to the corresponding elements of the same order in the Padé table for $\frac{F}{G}$.

Moreover, since the Padé table for t is normal, we can conclude that the Padé table for $\frac{F}{G}$ is normal for all orders $p \leqslant k$.

P.T. for t pseudo P.T. for t P.T. for $\frac{F}{G}$



(3.5) (3.6)

The upper triangles in both tables (3.5) and (3.6) are identical.

The above theorem and lemma give the possibility of considering $\frac{F}{G}$ as a Padé-like approximation, with about the same properties as the ordinary Padé approximations.

Furthermore, some of the results of the study of linear Runge-Kutta methods are not applicable on nonlinear methods anymore. An example of this is the restriction of the order in the linear case, proved by Butcher (1964). He proved that the highest attainable order with five function evaluations is four. But as mentioned by Scraton (1964), it is possible to derive a fifth order method with five function evaluations, using nonlinear combinations.

Let

$$k_1 = f(x_i, y_i)$$

$$k_2 = f(x_i + \tfrac{2}{9}h,\ y_i + \tfrac{2}{9}hk_1)$$

$$k_3 = f(x_i + \tfrac{1}{3}h,\ y_i + h(\tfrac{1}{22}k_1 + \tfrac{1}{4}k_2))$$

$$k_4 = f(x_i + \tfrac{3}{4}h,\ y_i + \tfrac{3h}{128}(23k_1 - 81k_2 + 90k_3))$$

$$k_5 = f(x_i + \tfrac{9}{10}h,\ y_i + \tfrac{9h}{10000}(-345k_1 + 2025k_2 - 1224k_3 + 544k_4))$$

and consider

$$y_{i+1} = y_i + hV + h\frac{US}{W}$$

with

$$V = \frac{17}{162}k_1 + \frac{81}{170}k_3 + \frac{32}{135}k_4 + \frac{250}{1377}k_5$$

$$U = -\frac{1}{18}k_1 + \frac{27}{170}k_3 - \frac{4}{15}k_4 + \frac{25}{153}k_5$$

$$S = \frac{19}{24}k_1 + \frac{27}{8}k_2 + \frac{57}{20}k_3 - \frac{4}{15}k_4$$

$$W = k_4 - k_1$$

This is proved to be an order five method.

Further investigations about the algebra concerned in deriving rational methods and their properties are in progress.

4. ACKNOWLEDGEMENT

I wish to thank L. Wuytack for his advice towards improvement of this paper.

5. LITERATURE

1. J. O. Lambert, B. Shaw, "On the numerical solution of $y' = f(x, y)$ by a class of formulae based on rational ap-

proximation", Math. Comp. vol. 19 # 91 July 1965, pp. 456-462.

2. J. C. Butcher, "On the attainable order of Runge-Kutta methods", Math. Comp. vol. 19 #91 July 1965, pp. 408-417.
3. R. E. Scraton, "Estimation of the truncation error in Runge-Kutta and allied processes", The Computer Journal vol. 17 #3 October 1964, pp. 246-248.
4. E. W. Cheney, "Introduction to approximation theory", McGraw-Hill, New York, 1966.
5. J. D. Lambert, "Computational methods in ordinary differential equations", John Wiley, London 1974.
6. P. Wynn, "On a device for computing the $e_m(S_n)$ transformation", MTAC vol. 10, 1956, p. 91.
7. P. Wynn, "Acceleration techniques for iterated vector and matrix problems", Math. Comp., vol. 16, 1962, p. 301-322.
8. H. Padé, "Sur la représentation approchée d'une fonction par des fractions rationnelles", Am. Sci. Ecole Normale Supér. (3) 9, 1892, pp. 1-92.
9. W. B. Gragg, "The Padé table and its relation to certain algorithms of numerical analysis", SIAM Review 14, 1972, pp. 1-62.